AD-A113 971    PURDUE UNIV LAFAYETTE IN DEPT OF STATISTICS    F/G 12/1
SELECTING PROCEDURES FOR OPTIMAL SUBSET OF REGRESSION VARIABLES--ETC(U)
JAN 82   S S GUPTA, D HUANG    N00014-75-C-0455
UNCLASSIFIED    MMS-82-2    NL

1 OF 1

END
DATE FILMED
5-82
DTIC

(12)



# PURDUE UNIVERSITY

## DEPARTMENT OF STATISTICS

## DIVISION OF MATHEMATICAL SCIENCES

12

# SELECTING PROCEDURES FOR OPTIMAL SUBSET OF REGRESSION VARIABLES*

by

Shanti S. Gupta, Purdue University
and
Deng-Yuan Huang, National Taiwan Normal University

Mimeograph Series #82-2

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #82-2

January 1982

*This research was supported by a grant from the National Science Council of Republic of China. It is also supported by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University.

# SELECTING PROCEDURES FOR OPTIMAL SUBSET OF REGRESSION VARIABLES*

by

Shanti S. Gupta, Purdue University
and
Deng-Yuan Huang, National Taiwan Normal University

Recently, a number of methods have been developed for selecting the "best" or at least a "good" subset of variables in regression analysis. For various reasons, we may be interested in including only a subset, say, of size $r < p$, the number of independent variables. Various authors have considered this problem and a variety of techniques are presently being used to construct such subsets.

Arvesen and McCabe (1975) proposed a procedure for selecting a subset within a class of subsets with t (fixed) independent variables, taking into account the statistical variation of the residual mean squares. Huang and Panchapakesan (1982) proposed a selection procedure based on the expected residual sums of squares. Hsu and Huang (1982) studied a sequential selection procedure for good regression models.

In this paper, we are interested in deriving an optimal decision procedure based on residual mean squares to select a subset excluding all "inferior" independent variables. This kind of optimality criterion is related to the approach of Gupta and Huang (1977).

Let $\pi_0, \pi_1, \ldots, \pi_k$ denote k+1 normal populations with unknown variances $\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2$ respectively. Assume that $\sigma_0^2$ is known. A population (model) is said to be superior (or good) if $\sigma_i^2 < \Delta\, \sigma_0^2$, to be inferior (or bad) if

*This research was supported by a grant from the National Science Council of Republic of China. It is also supported by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University.

$\sigma_i^2 \geq \Delta\, \sigma_0^2$, where $\Delta$ is a specified constant greater than 1. Let $\Omega$ be the parameter space which is the collection of all possible parameters.

Let CD stand for a correct decision which is defined to be the selection of any subset which excludes all the inferior populations.

Assuming the following model

$$(1) \qquad \underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$$

where $X = [\underset{\sim}{1}, \underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_{p-1}]$ is an nxp known matrix of rank $p \leq n$, $\underset{\sim}{\beta}' = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ is a 1xp parameter vector, and $\underset{\sim}{\epsilon} \sim N(\underset{\sim}{0}, \sigma_0^2 I_n)$, and $\underset{\sim}{1}' = [1, \ldots, 1]_{1xN}$, $I_n$ is an identity matrix with nxn.

In what follows, (1) which has p-1 independent variables, will be viewed as the true model. Without loss of generality we can assume that $\sigma_0^2 = 1$.

Consider the models for any r, $2 \leq r \leq p-1$,

$$(2) \qquad \underset{\sim}{Y} = X_{ri}\, \underset{\sim}{\beta}_{ri} + \underset{\sim}{\epsilon}_{ri}$$

where $X_{ri}$ is an nxr matrix of rank r with $X_{11}' = [1, \ldots, 1]_{1xn}$, $\underset{\sim}{\beta}_{ri}$ is a rx1 parameter vector, and $\underset{\sim}{\epsilon}_{ri} \sim N(\underset{\sim}{0}, \sigma_{ri}^2 I_n)$, $i = 1, 2, \ldots, k_r = \binom{p-1}{r-1}$. Let $k = \sum_{r=2}^{p-1} k_r$. It should be noted that in stating the reduced model (2), our comparisons of models are made under the true model assumptions. The goal is to include all the designs $X_{ri}$ (or sets of independent variables) associated with $\sigma_{[j]}^2$, $j = 1, \ldots, k-t$, where $\sigma_{[1]}^2 \leq \sigma_{[2]}^2 \leq \ldots \leq \sigma_{[k-t]}^2$ are ordered values from some of $\sigma_{ri}$'s, $i = 1, \ldots, k_r$, $r = 2, \ldots, p-1$.

Note that for any r, $2 \leq r \leq p-1$, if

$$SS_{ri} = \underset{\sim}{Y}'\{I - X_{ri}(X_{ri}'X_{ri})^{-1}X_{ri}'\}\underset{\sim}{Y} = \underset{\sim}{Y}'Q_i\underset{\sim}{Y},$$

then

$$SS_{ri} \sim \chi^2(\nu_r, (X\underset{\sim}{\beta})'Q_{ri}(X\underset{\sim}{\beta})/2)$$

(under the true model), where $\nu_r = n-r$, for $1 \le i \le k_r$. Note that the noncentrality parameter, in general, is not zero, and that

$$\sigma^2_{ri} = 1+(X\underset{\sim}{\beta})'Q_{ri}(X\underset{\sim}{\beta})/\nu_r.$$

Now we need some notation. Deleting a set of $\beta_i$'s without specifying which ones are deleted, we use ri to denote the special subset that is not deleted. For example, if p = 3, r = 2 then there are three subsets with size 2; namely, $\{\beta_1,\beta_2\}$, $\{\beta_1,\beta_3\}$ and $\{\beta_2,\beta_3\}$. Then r1 denotes the set $\{\beta_1,\beta_2\}$, r2 denotes $\{\beta_2,\beta_3\}$ and r3 denotes $\{\beta_1,\beta_3\}$. Then, we use $\tilde{\beta}$ to denote the vector with the following subsets: $\{\beta_1,\beta_2\}$ with $(\beta_1,\beta_2,0)$, $\{\beta_1,\beta_3\}$ with $(\beta_1,0,\beta_3)$, and $\{\beta_2,\beta_3\}$ with $(0,\beta_2,\beta_3)$, where 0 is the parameter value which is omitted from the true model of the appropriate $\beta_i$'s. Thus, in the following, we will use $\Omega_{0,ri}$ to denote those sets of $\underset{\sim}{\tilde{\beta}}$ as described above with the further condition that $\sigma^2_{ri} = \sigma^2_0 = 1$. Similarly, $\Omega_{1,ri}$ will be used to denote the sets of $\underset{\sim}{\tilde{\beta}}$ as described above with the further restriction that $\sigma^2_{ri} \ge \Delta$. Formally, we write

$$\Omega_{0,ri} = \{\underset{\sim}{\tilde{\beta}} \mid \sigma^2_{ri} = 1\},$$

and

$$\Omega_{1,ri} = \{\underset{\sim}{\tilde{\beta}} \mid \sigma^2_{ri} \ge \Delta\},$$

where $i = 1,\ldots,k_r$; $r = 2,\ldots,p-1$, and let

$$\Omega_1 = \bigcup_{r=2}^{p-1} \bigcup_{i=1}^{k_r} \Omega_{1,ri}, \text{ and}$$

$$\Omega_0 = \bigcap_{r=2}^{p-1} \bigcap_{i=1}^{k_r} \Omega_{0,ri}.$$

Let $g_{\sigma^2_{ri}}(s_{ri})$ denote the probability density of $S_{ri}$ depending on the parameter $\sigma^2_{ri}$, where $S_{ri} = \frac{SS_{ri}}{\nu_r}$, $i = 1,\ldots,k_r$; $r = 2,\ldots,p-1$.

Consider a family of hypotheses testing problems as follows:

$$\text{(3)} \qquad H_{0,ri}\colon\ \tilde{\underline{\beta}} \in \Omega_0 \quad \text{vs} \quad K_{ri}\colon\ \tilde{\underline{\beta}} \in \Omega_{1,ri};$$

$i = 1,\ldots,p-1$, $r = 2,\ldots,p-1$. A test of the hypotheses (3) will be defined to be a vector $(\varphi_1(\underline{y}),\ldots,\varphi_k(\underline{y}))$, where the elements of the vector are ordinary test functions; when $\underline{y}$ is observed we reject $H_{0,t}$ with probability $\varphi_t(\underline{y})$, $1 \le t \le k$. The power function of a test $(\varphi_1,\ldots,\varphi_k)$ is defined to be the vector $(p_1(\tilde{\underline{\beta}}),\ldots,p_k(\tilde{\underline{\beta}}))$ where

$$p_t(\tilde{\underline{\beta}}) = E_{\tilde{\underline{\beta}}}\, \varphi_t(\underline{Y}),$$

$1 \le t \le k$. Let $S(\gamma)$ be the set of all tests $(\varphi_1,\ldots,\varphi_k)$ such that

$$\text{(4)} \qquad E_{\tilde{\underline{\beta}}}\, \varphi_t(\underline{Y}) \le \gamma, \quad \tilde{\underline{\beta}} \in \Omega_0.$$

We define $\varphi^0 = (\varphi^0_1,\ldots,\varphi^0_k)$ as

$$\varphi^0_{ri}(\underline{y}) = \begin{cases} 1, & \text{if } g_\Delta(s_{ri}) \ge c\, g_1(s_{ri}), \\ 0, & \text{if } g_\Delta(s_{ri}) < c\, g_1(s_{ri}), \end{cases}$$

such that $E_{\tilde{\underline{\beta}}}\, \varphi^0_{ri}(\underline{Y}) = \gamma$, $\tilde{\underline{\beta}} \in \Omega_0$, where $s_{ri}$ is the observed value of $S_{ri}$. It can be shown that $\varphi^0$ maximizes

$$\min_{1 \le t \le k} \inf_{\tilde{\beta} \in \Omega_{1,t}} E_{\tilde{\beta}}\, \varphi_t(\underline{Y})$$

among all tests $\varphi = (\varphi_1, \ldots, \varphi_k) \in S(\gamma)$ (cf. Gupta and Huang (1977)).

To determine the constant c, we proceed follows: for a given $\eta > 0$, there exists a smallest positive integer $k_0$ such that

$$\frac{a_{k_0}}{\eta} < 1 \quad \text{and} \quad \frac{a_{k_0+1}}{a_{k_0}} + \frac{a_{k_0}}{\eta} \le 1,$$

where

$$a_\ell(s_{ri}) = \frac{e^{-\lambda_r} \lambda_r^\ell}{\ell!} \left[\frac{\nu_r s_{ri}}{2}\right]^\ell \frac{\Gamma(\frac{1}{2}\nu_r)}{\Gamma(\frac{1}{2}\nu_r + \ell)},$$

$\ell = 0, 1, 2, \ldots$; $\lambda_r = \frac{(\Delta - 1)\nu_r}{2}$. For this $k_0$, it can be shown that

$$0 < \frac{g_\Delta(s_{ri})}{g_1(s_{ri})} - \sum_{\ell=0}^{k_0-1} a_\ell(s_{ri}) = \sum_{k=0}^{\infty} a_{k_0+k} \le \eta,$$

where

$$\frac{g_\Delta(s_{ri})}{g_1(s_{ri})} = \sum_{\ell=0}^{\infty} a_\ell .$$

Thus, approximately,

$$\frac{g_\Delta(s_{ri})}{g_1(s_{ri})} \approx \sum_{\ell=0}^{k-1} a_\ell(s_{ri})$$

with error less than $\eta$. For $\tilde{\beta} \in \Omega_0$,

$$E_{\underset{\sim}{\beta}}\, \varphi^{0}(\underset{\sim}{Y}) = P_{\underset{\sim}{\beta}}\{g_\Delta(S_{ri}) \geq c\; g_1(S_{ri})\}$$

$$= P_{\underset{\sim}{\beta}}\{\sum_{\ell=0}^{k_0-1} a_\ell(S_{ri}) \geq c\}$$

$$= \int_0^\infty I_{[\sum_{\ell=0}^{k_0-1} a_\ell(s_{ri}) \geq c]}(s_{ri})\, g_1(s_{ri})\, ds_{ri} = \gamma,$$

where $g_1(s_{ri})$ is the central $\chi^2$ with $\nu_r$ degrees of freedom and $I_A(x) = 0$ for $x \notin A$, $I_A(x) = 1$ for $x \in A$. The constant c can be determined.

## References

[1] Arvesen, J. N. and McCabe, G. P. Jr. (1975). Subset selection problems for variances with applications to regression analysis. JASA, 70, 166-170.

[2] Gupta, S. S. and Huang, D. Y. (1977). Some multiple decision problems in analysis of variance. Comm. Statist. A-Theory Methods, 6, 1035-1054.

[3] Hsu, T. A. and Huang, D. Y. (1982). Some sequential selection procedures for good regression models. Comm. Statist. A-Theory Methods, 11, 411-421.

[4] Huang, D. Y. and Panchapakesan, S. (1982). On eliminating inferior regression models. To appear in Comm. Statist.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>Technical Report #82-2 | 2. GOVT ACCESSION NO.<br>AD-A113971 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>SELECTING PROCEDURES FOR OPTIMAL SUBSET OF REGRESSION VARIABLES | | 5. TYPE OF REPORT & PERIOD COVERED<br>Technical Report |
| | | 6. PERFORMING ORG. REPORT NUMBER<br>Technical Report #82-2 |
| 7. AUTHOR(s)<br>Shanti S. Gupta and Deng-Yuan Huang | | 8. CONTRACT OR GRANT NUMBER(s)<br>N00014-75-C-0455 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Purdue University<br>Department of Statistics<br>West Lafayette, IN 47907 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>042-243 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Office of Naval Research<br>Washington, DC | | 12. REPORT DATE<br>January 1982 |
| | | 13. NUMBER OF PAGES<br>7 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) | | 15. SECURITY CLASS. (of this report) |
| | | 15a. DECLASSIFICATION DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Selection Procedures, Regression Variables, Optimal Subset

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Recently, a number of methods have been developed for selecting the "best" or at least a "good" subset of variables in regression analysis. For various reasons, we may be interested in including only a subset, say, of size $r < p$, the number of independent variables. Various authors have considered this problem and a variety of techniques are presently being used to construct such subsets. In this paper, we are interested in deriving an optimal decision procedure based on residual mean squares to select a subset excluding all "inferior" independent variables. This kind of optimality criterion is related to the approach of Gupta and Huang (1977).

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

82